

CG020 Genomika

Přednáška 1

Úvod do bioinformatiky

Jan Hejátko

Funkční genomika a proteomika rostlin,
Středoevropský technologický institut (CEITEC)
a

Národní centrum pro výzkum biomolekul,
Přírodovědecká fakulta,

Masarykova univerzita, Brno
hejatko@sci.muni.cz, www.ceitec.eu

MUNI
SCI



Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
 - Spektrum „on-line“ zdrojů
 - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
 - GENOMOVÉ zdroje
- Analytické nástroje
 - Vyhledávání homologií
 - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restrikčních míst...
 - Další [www genomové nástroje](#)

Schéma předmětu

- **Kapitola 01**
 - **Úvod do bioinformatiky**
- **Kapitola 02**
 - **Identifikace genů**
- **Kapitola 03**
 - **Přístupy reverzní genetiky**
- **Kapitola 04**
 - **Přístupy genetiky přímé**

Schéma předmětu

- **Kapitola 05**
 - **RNA interference a genomové editování**
- **Kapitola 06**
 - **Genová exprese a chemická genetika**
- **Kapitola 07**
 - **Protein-proteinové interakce a jejich analýza**
- **Kapitola 08**
 - **Současné metody sekvenování DNA**

Schéma předmětu

- **Kapitola 09**
 - **Struktura genomů**
- **Kapitola 10**
 - **Evoluce genomů**
- **Kapitola 11**
 - **Genomika a systémová biologie**
- **Kapitola 12**
 - **Praktické aspekty funkční genomiky**
 - **Modelové organismy**
 - **PCR**

Literatura

- Literární zdroje pro kapitulu 01:
 - **Bioinformatics and Functional Genomics**, 3rd Edition, Jonathan Pevsner, Wiley-Blackwell, 2015
<http://www.bioinfbook.org/php/?q=book3>
 - **Úvod do praktické bioinformatiky**, Fatima Cvrčková, 2006, Academia, Praha
 - **Plant Functional Genomics**, ed. Erich Grotewold, 2003, Humana Press, Totowa, New Jersey

Osnova

- Schéma předmětu
- Definice

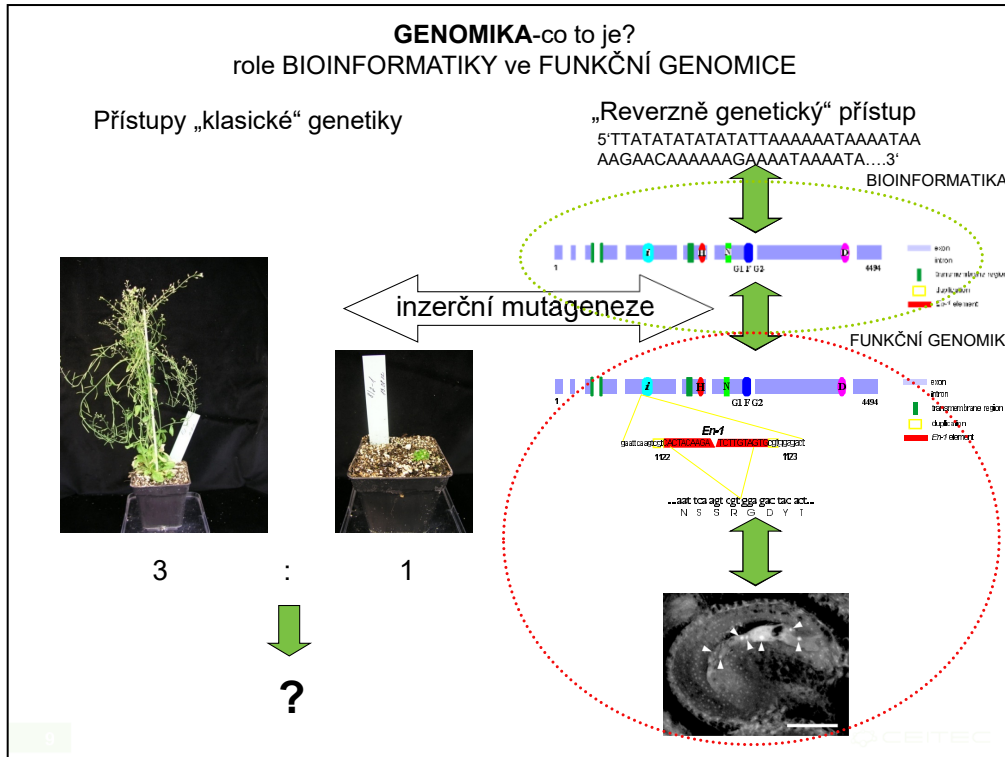
GENOMIKA-co to je?

- *Sensu lato* (v širším pojetí) zkoumá **STRUKTURU** a **FUNKCI genomů**
 - Předpokladem je znalost genomu (sekvencí)-práce s databázemi
- *Sensu stricto* (v užším pojetí) zkoumá **FUNKCI jednotlivých genů** - **FUNKČNÍ GENOMIKA**
 - používá zejména přístupy **REVERZNÍ GENETIKY**

Genomics is a science discipline that is interested in the analysis of genomes. Genome of each organism is a complex of all genes of the respective organism. The genes could be located in cytoplasm (prokaryots) nucleus (in most eukaryotic organisms), mitochondria or chloroplasts (in plants).

The critical prerequisite of genomics is the knowledge of gene sequences.

Functional genomics is interested in function of individual genes.



With the knowledge of gene sequences (or the knowledge of the gene files in the individual organisms, i.e. the knowledge of genomes), **Reverse Genetics** appears that allows study their function.

In comparison to "classical" or **Forward Genetics**, starting with the phenotype, the reverse genetics starts with the sequence identified as a gene in the sequenced genome. The gene identification using approaches of **Bioinformatics** will be described later (see Lesson 02).

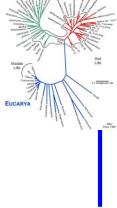
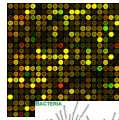
Reverse genetics uses a spectrum of approaches that will be described in the Lesson 03 that allow isolation of sequence-specific mutants and thus their phenotype analysis.

The necessity of having phenotype alterations in the forward genomics approach introduces important difference between those two approaches. Thus, the gene is no longer understood as a factor (*trait*) determining *phenotype*, but rather as a piece of DNA characterized by the unique *string of nucleotides*. i.e. **physical DNA molecule**.

Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY

Bioinformatika



- **Definice bioinformatiky** (podle NIH vědeckého a technologického konsorcia pro biomedicínské informace)

Výzkum, vývoj nebo aplikace výpočetních nástrojů a přístupů za účelem zvyšování rozvoje využití biologických, lékařských, dat o chování nebo zdraví, včetně těch, které umožňují taková data získávat, ukládat, organizovat, archivovat, analyzovat nebo vizualizovat.

11

CEITEC

NIH WORKING DEFINITION OF BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

July 17, 2000

The following working definition of bioinformatics and computational biology were developed by the BISTIC Definition Committee and released on July 17, 2000. The committee was chaired by Dr. Michael Huerta of the National Institute of Mental Health and consisted of the following members:

Bioinformatics Definition Committee BISTIC Members Expert Members

Michael Huerta (Chair) Gregory Downing
Florence Haseltine Belinda Seto
Yuan Liu

Preamble

Bioinformatics and computational biology are rooted in life sciences as well as computer and information sciences and technologies. Both of these interdisciplinary approaches draw from specific disciplines such as mathematics, physics, computer science and engineering, biology, and behavioral science. Bioinformatics and computational biology each maintain close interactions with life sciences to realize their full potential. Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology. Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface.

Definition

The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

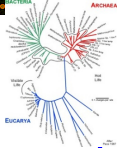
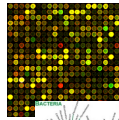
Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

What is Bioinformatics?

- Interface of **biology** and **computers**
- Analysis of **proteins, genes** and **genomes** using **computer algorithms** and **computer databases**
- **Genomics** is the **analysis of genomes**. The **tools of bioinformatics** are used **to make sense** of the **billions of base pairs of DNA** that are sequenced by genomics projects.

J. Pevsner,
<http://www.bioinfbook.org/index.php>

Bioinformatika



- **Bioinformatika ve funkční genomice**
 - **Zpracování a analýza sekvenačních dat**
 - Identifikace referenčních sekvencí
 - Identifikace genů
 - Identifikace homologů, ortologů a paralogů
 - Korelační analýzy mezi genomy a fenotypy (včetně člověka)
 - **Zpracování a analýza transkripčních dat**
 - Transkripční profilování pomocí DNA čipů nebo next-gen sekvenování
 - **Vyhodnocování experimentálních dat a predikce nových regulací** v přístupech **systémové biologie**
 - Matematické modelování genových regulačních sítí

Množství informace v DNA

- Every **bp**= **4 bits**
- Human genome = **~3 billion bp**
 - = $4 \times 3 \times 10^9$
 - = 1.2×10^{10} bits
 - = 1.5×10^9 bytes (**1.5 GB**)
- This amount of information is contained in a cell nucleus with **10 μ m** diameter

- There is **~2m** of DNA in **every somatic human cell**
 - **Each human** is composed of about **10¹² cells**
 - Thus every human contains 2×10^{12} of DNA
 - = **2 \times 10⁹km** of DNA
 - Distance from the sun to Uranus = 2.8×10^9 km
 - **Each single human contains enough DNA to stretch from the sun to Uranus**

Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
 - Spektrum „on-line“ zdrojů

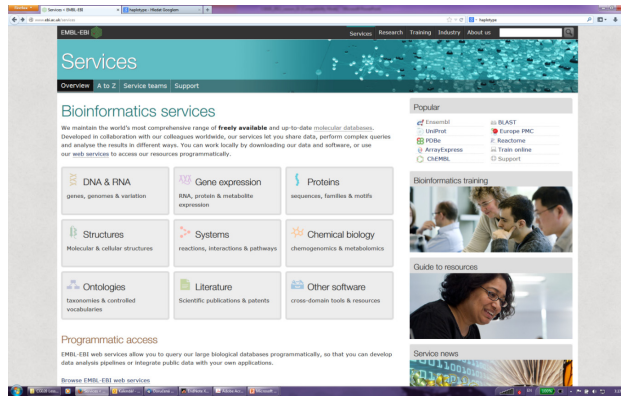
Spektrum on-line zdrojů

EMBLnet National Nodes		
Vienna BioCenter	Austria	http://www.at.emblnet.org/
EMBL	Belgium	http://www.be.emblnet.org/
EBI/EMBL	Denmark	http://dbi-base.dk/
CSC	Finland	http://www.fi.emblnet.org/
INFORMAGEN	France	http://www.infololgen.fr/
GEMISnet	Germany	http://gemisnet.dfbp-halleberg.de/biocont/
IMSB	Greece	http://www.imsb.forth.gr/
EMBL	Hungary	http://www.hu.emblnet.org/
INCE	Ireland	http://www.gps.tcd.ie/
EMBL	Israel	http://dgpas.welmann.ac.il/foef/fin.html
EMBL-ABR	Italy	http://dbi-www.ba.cnr.it/2000/BioWWW/Bo-WWW.htm
KAOS/CANN	Netherlands	http://www.kaos.kun.nl/
EMBL	Norway	http://www.no.emblnet.org/
EMBL	Poland	http://www.ils.waw.pl/
EMBL	Portugal	http://www.lgc.gubernet.pt/
GEMISnet	Russia	http://www.gemisnet.msk.ru/
CNB-CSC	Spain	http://www.es.emblnet.org/
EMBL	Sweden	http://www.se.emblnet.org/
EMBL	Switzerland	http://www.ch.emblnet.org/
EMBLNET	UK	http://www.seqnet.rl.ac.uk/
EMBLnet Specialist Nodes		
EMBL	Germany	http://www.mips.biochem.mpg.de/
EMBL	Italy	http://www.ligabiochemia.it/
Pharmacia Upjohn	Sweden	http://www.gnu.com/
F. Hoffmann La Roche	Switzerland	http://www.niche.com/
EMBL	UK	http://www.abi.ac.uk/
EMBL-RC	UK	http://www.bgmp.mrc.ac.uk/
Sanger	UK	http://www.sanger.ac.uk/
EMBLER	UK	http://www.bioinf.man.ac.uk/ebrewer
EMBLnet Associate Nodes		
EMBL	Argentina	http://fsl.biol.unla.edu.ar/emblnet
EMBL	Australia	http://www.emph.usc.edu.au/
EMBL	China	http://www.cbi.pku.edu.cn/
EMBL	Cuba	http://dbi.cigb.edu.cu/
EMBL	India	http://sankarj.emblnet.org.in/
EMBL	South Africa	http://www.sabi.ac.za
EMBL Information Providers		
EMBL	USA	http://www.ncbi.nlm.nih.gov/
EMBL	USA	http://www.nlm.nih.gov/
EMBL	USA	http://www.nih.gov/

There are many of on-line resources that could be used.

Spektrum on-line zdrojů

- EBI <http://www.ebi.ac.uk/services>



17

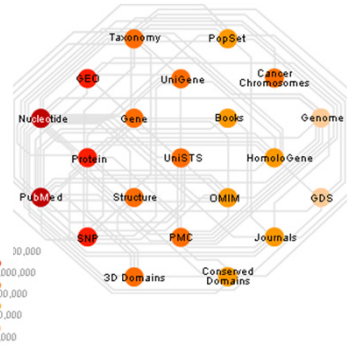
CEITEC

Nowadays, the resources are interconnected and could be accessed via dedicated web pages. Among the best and mostly used www resources integrating plenty of database resources belong www portal of European Bioinformatics Institute (EBI) in Europe (Germany) and National Center of Biotechnology Information (NCBI) in the USA (

Spektrum on-line zdrojů

NCBI <http://www.ncbi.nlm.nih.gov/>

The screenshot shows the NCBI homepage with a search bar and navigation menu. The main content area includes a 'Welcome to NCBI' message, a 'Get Started' section with links to tools and tutorials, a 'NCBI YouTube channel' promotion, and a 'Popular Resources' list containing items like PubMed, Bookshelf, and BLAST.



18

CEITEC

Nowadays, the resources are interconnected and could be accessed via dedicated web pages.

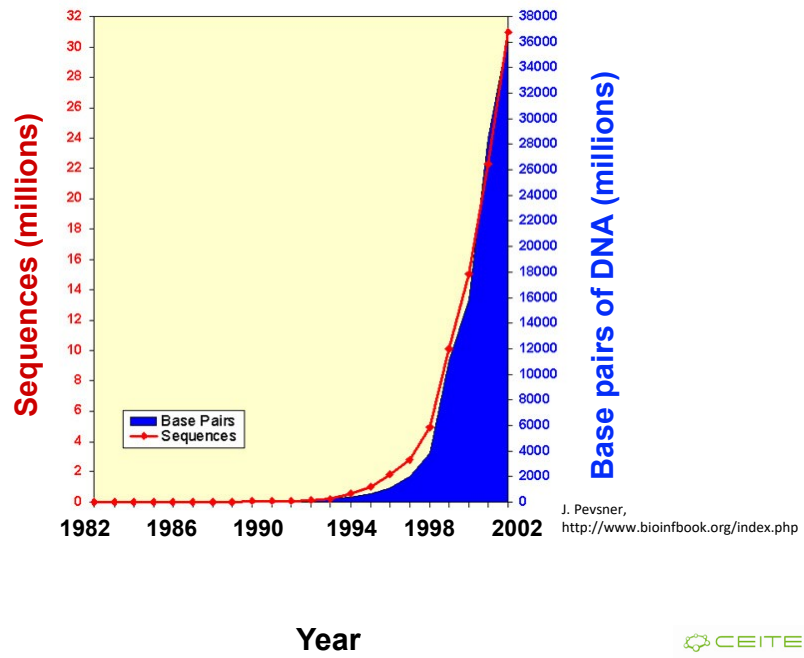
Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
 - Spektrum „on-line“ zdrojů
 - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze

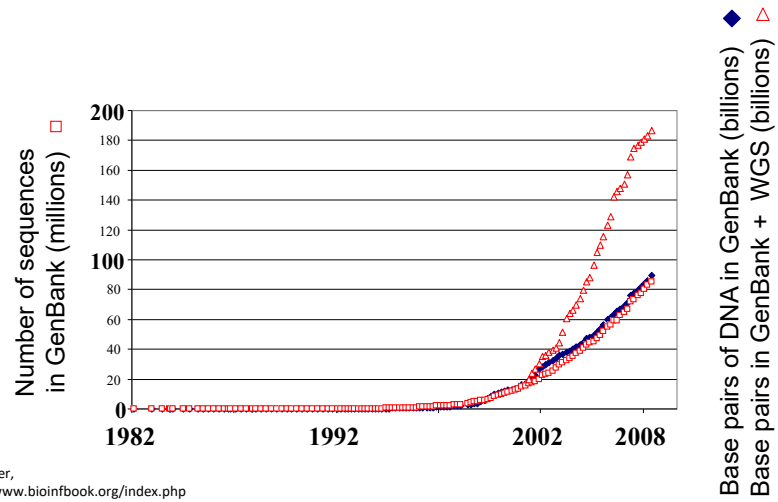
Primární databáze

- zahrnují soubory primárních dat – sekvencí DNA a proteinů
 - Sekvence v databázích tzv. „Velké trojky“:
 - **EMBL**
 - <http://www.ebi.ac.uk/embl/>
 - **GenBank,**
 - <https://www.ncbi.nlm.nih.gov/>
 - **DDBJ,**
 - <http://www.ddbj.nig.ac.jp>
 - denně vzájemná výměna a zálohování dat
 - velká datová náročnost (kapacita i software)

Growth of GenBank

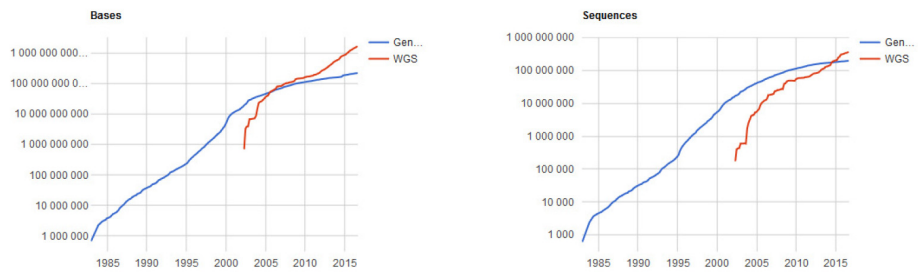


Growth of GenBank + Whole Genome Shotgun (1982-November 2008): we reached 0.2 terabases

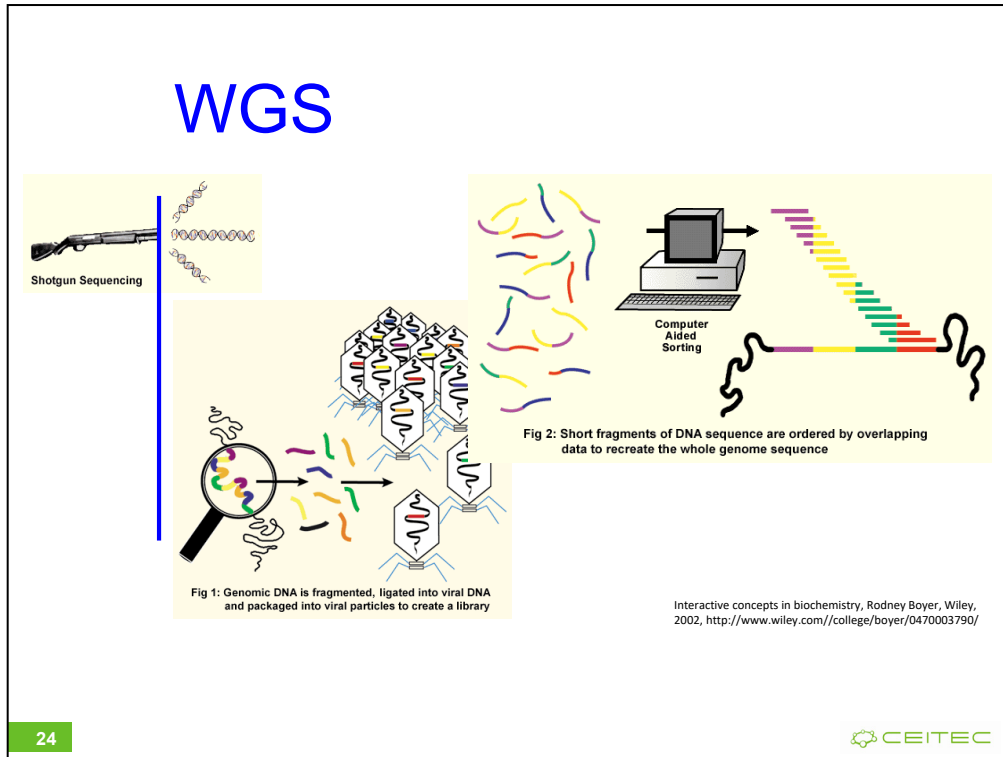


J. Pevsner,
<http://www.bioinfbook.org/index.php>

Growth of GenBank Aug 2016



- Prosinec **1982** 680 338 bp, 606 sekvencí
- Duben **2002** 19×10^9 bp, 17×10^6 sekvencí + WGS 692×10^6 bp, 172 768 sekvencí
- Srpen **2016** 218×10^9 bp, 196×10^6 sekvencí + WGS $1,6 \cdot 10^{12}$ bp, 360×10^6 sekvencí

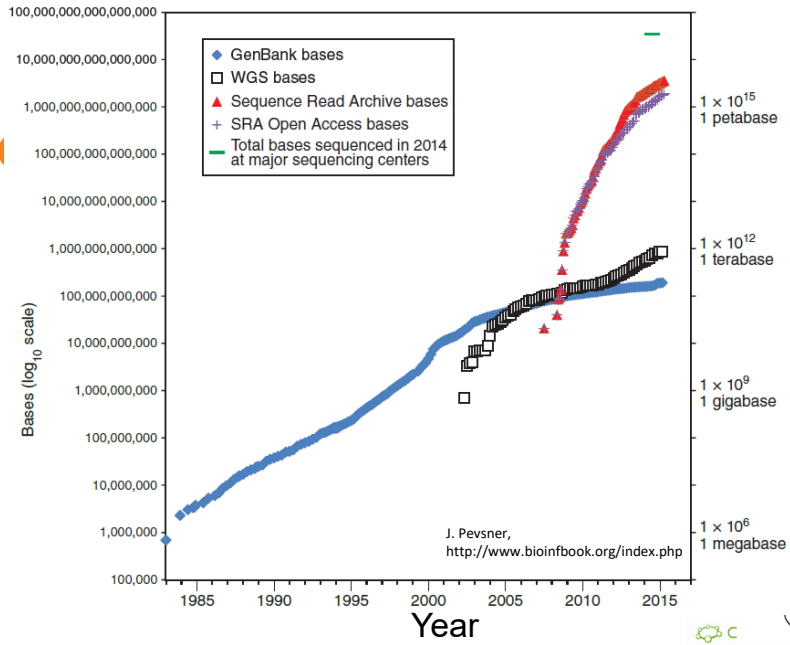


Shotgun sequencing allows a scientist to rapidly determine the sequence of very long stretches of DNA. The key to this process is fragmenting of the genome into smaller pieces that are then sequenced side by side, rather than trying to read the entire genome in order from beginning to end. The genomic DNA is usually first divided into its individual chromosomes. Each chromosome is then randomly broken into small strands of hundreds to several thousand base pairs, usually accomplished by mechanical shearing of the purified genetic material. Each of the short DNA pieces is then inserted into a DNA vector (a viral genome), resulting in a viral particle containing "cloned" genomic DNA (Fig. 1).

The collection of all the viral particles with all the different genomic DNA pieces is referred to as a library. Just as a library consists of a set of books that together make up all of human knowledge, a genomic library consists of a set of DNA pieces that together make up the entire genome sequence. Placing the genomic DNA within the viral genome allows bacteria infected with the virus to faithfully replicate the genomic DNA pieces. Additionally, since a little bit of known sequence is needed to start the sequencing reaction, the reaction can be primed off the known flanking viral DNA.

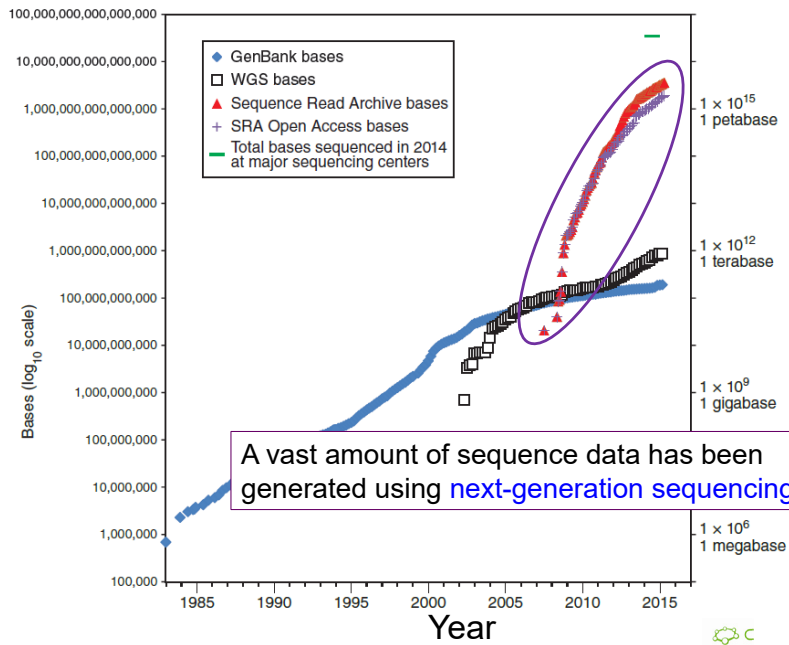
In order to read all the nucleotides of one organism, millions of individual clones are sequenced. The data is sorted by computer, which compares the sequences of all the small DNA pieces at once (in a "shotgun" approach) and places them in order by virtue of their overlapping sequences to generate the full-length sequence of the genome (Fig. 2). To statistically ensure that the whole genome sequence is acquired by this method, an amount of DNA equal to five to ten times the length of the genome must be sequenced. (Interactive concepts in biochemistry, Rodney Boyer, Wiley, 2002, <http://www.wiley.com/college/boyer/0470003790/>)

Growth of DNA Sequence in Repositories

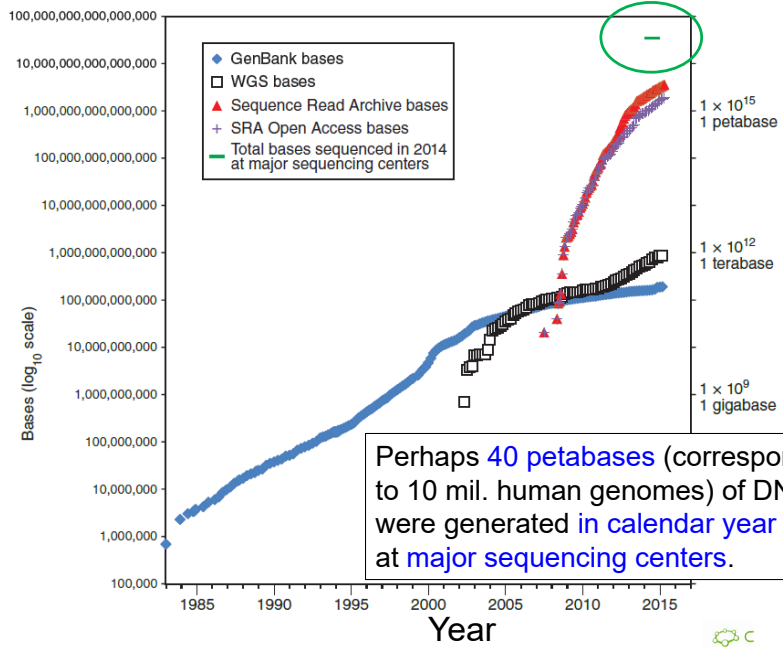


33

Growth of DNA Sequence in Repositories



Growth of DNA Sequence in Repositories



B&FG 3e
Fig. 2-3

27 22



Primární databáze

- zahrnují soubory primárních dat – sekvencí DNA a proteinů
 - **Proteinové sekvence:**
 - **PIR**, <http://pir.georgetown.edu/>
 - **MIPS**, <http://www.mips.biochem.mpg.de>
 - **SWISS-PROT**, <http://www.expasy.org/sprot/>

Primární databáze

- Typy sekvencí v primárních databázích
 - Standardní nukleotidové sekvence získané kvalitním sekvencováním
 - **ESTs** (Expressed Sequence Tags)
 - **HGTS** (High Throughput Genome Sequencing)
 - neanotované „surové“ výsledky sekvenačních projektů
 - Referenční sekvence anotovaných genomů
 - **TPAs** (Third Party Annotation)
 - sekvence anotované jinými než původními autory

Primární databáze

GenBank (NCBI) <https://www.ncbi.nlm.nih.gov/>



The screenshot shows the NCBI homepage with a search bar at the top, a navigation menu on the left, and a main content area with sections for 'Welcome to NCBI', 'Get Started', 'NCBI YouTube channel', 'Popular Resources', and 'NCBI Announcements'.

NCBI Home

- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Get Started

- Tools:** Analyze data using NCBI software
- Downloads:** Get NCBI data or software
- BioLinks:** Learn how to accomplish specific tasks at NCBI
- Submissions:** Submit data to GenBank or other NCBI databases

NCBI YouTube channel

Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel.

GO

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI Announcements

New version of Gen... available

An integrated, downlo... for viewing and analy...

NCBI's July Newslett...

Primární databáze

Gene symbol: *adh*
Gene description: non-component VMA-like sensor kinase
Gene tag: *adh*
Gene type: protein coding
Reference(s): PMID1976244
Organism: *Agrobacterium tumefaciens* subsp. *Agrobacterium tumefaciens* biovar. *Agrobacterium tumefaciens*
Taxonomy: Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Rhizobiales; Rhizobiales/Agrobacterium group; Agrobacterium; Agrobacterium tumefaciens complex

Genomic context
Location: chr1:141,484..146,163
Sequence: NC_022777.1 (141,484..146,163)

Genomic regions, transcripts, and products
Genomic Sequence: NC_022777.1

Gene ID: 819972.1
Gene Name: NC_022777.1 (141,484..146,163)
Gene type: protein coding
Gene length: 579
Gene structure: 5' UTR, Exon 1, Intron 1, Exon 2, 3' UTR
Gene model: NC_022777.1 (141,484..146,163)
Gene model: NC_022777.1 (141,484..146,163)
Gene model: NC_022777.1 (141,484..146,163)

Related articles
1. Sequence analysis of the *adh* gene from *Agrobacterium tumefaciens* strain T-3002. *Biochimica et Biophysica Acta* 1994;124:1-10.
2. The *adh* gene is a homolog of the *adh* gene from *Agrobacterium tumefaciens*. *Plant Cell* 1993;5:1105-1110.
3. Characterization of the *adh* gene of *Agrobacterium tumefaciens*. A transcriptional regulator and host gene ortholog. *Genes & Dev* 1997;11:197-209.
4. Analysis of the complete nucleotide sequence of the *adh* gene from *Agrobacterium tumefaciens* strain T-3002. *Plant Cell* 1993;5:1105-1110.

GeneID's Gene Reference Use Function: [What's New?](#)

Submit: [View Details](#) [Citations](#)

Primární databáze

The screenshot displays a GenBank entry for the gene **NP_059797.1**. The entry details include:

- Gene: **NP_059797.1**
- NP_059797.1: two-component VirA-like sensor kinase
- total range: NC_002377.1 (145,694..148,183)
- total length: 2,490
- strand: plus
- protein product length: 829

Links & Tools

- GenBank View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1](#)
- FASTA View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1](#)
- BLAST Genomic: [NC_002377.1 \(145,694..148,183\)](#)
- Graphical View: [NP_059797.1](#)
- BLAST Protein: [NP_059797.1](#)
- BLINK Results: [NP_059797.1](#)

The interface also shows a **Bibliography** section and **Related articles in PubMed**.

Primární databáze

The screenshot displays the NCBI Nucleotide database entry for the plasmid Ti. The accession number **NC_002237** is circled in red and labeled "Přístupový kód". The GeneBank identifier **U00000** is also circled in red and labeled "GeneBank Identifier". The entry details include:

- LOCUS:** NC_002237 2490 bp DNA linear MT 19-DEC-2003
- DEFINITION:** Agrobacterium tumefaciens octopine plasmid Ti, complete sequence.
- ACCESSION:** NC_002237
- VERSION:** NC_002237.1
- KEYWORDS:** Agrobacterium tumefaciens (Rhizobium radiobacter)
- SOURCE:** Agrobacterium tumefaciens (Strain: Schramm 1964) Schramm, R., Hooykaas, P.J., Parand, E.K., and Witsens, S.C.
- TITLE:** Octopine-type Ti plasmid sequence
- COMMENT:** Submitted (07-MAR-2004) Microbiology, Cornell University, Wing Hall, Ithaca, NY 14853, USA
- FEATURES:** Includes a **source** feature with coordinates 1..2490 and a **gene** feature for **virA** with coordinates 1..2490.

What is an **Accession Number**?

An **accession number** is **label** that **used to identify a sequence**. It is a **string of letters and/or numbers** that corresponds to a **molecular sequence**.

Examples (all for retinol-binding protein, RBP4):

X02775	GenBank genomic DNA sequence	DNA
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	

N91759.1	An expressed sequence tag (1 of 170)	RNA
NM_006744	RefSeq DNA sequence (from a transcript)	

NP_007635	RefSeq protein	Protein
AAC02945	GenBank protein	
Q28369	SwissProt protein	
1KT7	Protein Data Bank structure record	

J. Pevsner,
<http://www.bioinfbook.org/index.php>

NCBI's important **RefSeq** project: best **representative sequences**

RefSeq (accessible via the main page of NCBI) provides an **expertly curated accession number** that corresponds to **the most stable, agreed-upon "reference" version of a sequence**.

RefSeq identifiers include the following formats:

Complete genome	NC_#####
Complete chromosome	NC_#####
Genomic contig	NT_#####
mRNA (DNA format)	NM_##### e.g. NM_006744
Protein	NP_##### e.g. NP_006735

J. Pevsner,
<http://www.bioinfbook.org/index.php>

RefSeq

two-component VfrA-like sensor kinase

NCBI Reference Sequences (RefSeq)

Genome Annotation

The following sections contain reference sequences that belong to a specific genome build. [Explain](#)

Reference assembly

Genomic

1. **NC_003065.3**

Range: 190831..193332
Download: [GenBank](#), [FASTA](#), [Sequence Viewer \(Graphics\)](#)

mRNA and Protein(s)

1. **NP_396486.1** two component sensor kinase [Agrobacterium tumefaciens str. C58]

UniProtKB/Swiss-Prot: [P18540](#)

Conserved Domains (3) [summary](#)

cd08025	HATPase_c, Histidine kinase-like ATPase. This family includes several ATP-binding proteins for example: histidine kinase, DNA gyrase B, topoisomerase, heat shock protein HSP90, phytochrome-like ATPases and DNA mismatch repair proteins.
cd08032	HtkA, Histidine Kinase A (dimerization/phosphoreceptor) domain; Histidine Kinase. Location:468 - 530. A dimer is formed through parallel association of 2 domains creating 4-helix bundles; usually these domains contain a conserved His residue and are activated via...
P06138	PRK13837: two-component VfrA-like sensor kinase; Provisional. Location:18 - 833. Blast Score: 2944.

Related Sequences

NCBI's RefSeq project: many accession number formats for genomic, mRNA, protein sequences

Accession	Molecule	Method	Note
AC_123456	Genomic	Mixed	Alternate complete genomic
AP_123456	Protein	Mixed	Protein products; alternate
NC_123456	Genomic	Mixed	Complete genomic molecules
NG_123456	Genomic	Mixed	Incomplete genomic regions
NM_123456	mRNA	Mixed	Transcript products; mRNA
NM_123456789	mRNA	Mixed	Transcript products; 9-digit
NP_123456	Protein	Mixed	Protein products;
NP_123456789	Protein	Curation	Protein products; 9-digit
NR_123456	RNA	Mixed	Non-coding transcripts
NT_123456	Genomic	Automated	Genomic assemblies
NW_123456	Genomic	Automated	Genomic assemblies
NZ_ABCD12345678	Genomic	Automated	Whole genome shotgun data
XM_123456	mRNA	Automated	Transcript products
XP_123456	Protein	Automated	Protein products
XR_123456	RNA	Automated	Transcript products
YP_123456	Protein	Auto. & Curated	Protein products
ZP_12345678	Protein	Automated	Protein products

J. Pevsner,
<http://www.bioinfbook.org/index.php>

Primární databáze

The screenshot displays a GenBank entry for the gene **NP_059797.1**. The entry is located on the chromosome **NC_002377.1**, which is 2.9 Kbp long. The gene's total range is from 145,694 to 148,183. The gene is 2,490 nucleotides long and is transcribed on the plus strand. The protein product is 829 amino acids long. The entry includes a "Links & Tools" section with the following links: GenBank View, FASTA View, BLAST Genomic, Graphical View, BLAST Protein, and BLINK Results. A green arrow points to the gene name in the "Genes" track. Below the "Links & Tools" section, there are sections for "Bibliography" and "Related articles in PubMed".

NC_002377.1: 145K..148K (2.9Kbp)

Genes

NP_059797.1

NP_059797.1: two-component VirA-like sensor kinase
total range: NC_002377.1 (145,694..148,183)
total length: 2,490
strand: plus
protein product length: 829

Links & Tools

GenBank View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1 \(145,694..148,183\)](#)
FASTA View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1 \(145,694..148,183\)](#)
BLAST Genomic: [NC_002377.1 \(145,694..148,183\)](#)
Graphical View: [NP_059797.1](#)
BLAST Protein: [NP_059797.1](#)
BLINK Results: [NP_059797.1](#)

Bibliography

Related articles in PubMed

Primární databáze

The screenshot shows the NCBI GenBank entry for Agrobacterium tumefaciens plasmid Ti. The main content is the DNA sequence, which is displayed in a monospaced font. The sequence is preceded by a header: "Agrobacterium tumefaciens plasmid Ti, complete sequence" and "NCBI Reference Sequence: NC_002377.1". To the right of the sequence, there are several interactive panels: "Change region shown" (with a range of 145694 to 148183), "Customize view", "Analyze this sequence" (with options for BLAST, Primers, and Features), "Related information" (with links for Protein, Clusters, PubMed, etc.), and "Recent activity". The browser's address bar shows the URL "https://www.ncbi.nlm.nih.gov/nuccore/NC_002377.1".

Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáváním primárních dat (sekvencí)
- PROSITE, <http://www.expasy.org/prosite/>

The screenshot shows the ScanProsite web interface. At the top, there is a navigation bar with links: Home page, Site Map, Search ExPASy, Contact us, Swiss-Prot, PROSITE, and Proteomics tools. Below this is a search bar with the text "Search PROSITE" and a "Go" button. The main heading is "prosite ScanProsite". A brief description states: "This program allows to scan a protein sequence (either from Swiss-Prot or TrEMBL, or provided by the user) for the occurrence of patterns and profiles stored in the PROSITE database, or to search protein databases with a user-entered pattern." Below this are three bullet points: "enter a PROSITE accession number or pattern to search the Swiss-Prot/TrEMBL, and/or PDB databases with a pattern, OR", "enter a sequence or a Swiss-Prot/TrEMBL accession number to scan the sequence with all patterns, profiles and rules in PROSITE, OR", and "fill in both fields to find all occurrences of a pattern or profile in a sequence." The interface is divided into two main sections: "Scan a protein for PROSITE matches" and "Search Swiss-Prot with a PROSITE entry". The left section contains a text input field for a protein sequence, a "Clear" button, and a "START THE SCAN" button. The right section contains a text input field for a PROSITE accession number or pattern, a "Clear" button, and a "START THE SCAN" button. There are also checkboxes for "and specify which motifs to use" and "Advanced options".

Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáváním primárních dat (sekvencí)
- PROSITE, <http://www.expasy.org/prosite/>

```
>PDOC00003 PS00003 SULFATION Tyrosine sulfation site [rule] [Warning: rule with a high probability of occurrence].
171 - 585 sbesantYetsann

>PDOC00004 PS00004 CAMP_PHOSPHO_SITE cAMP- and cGMP-dependent protein kinase phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].
744 - 747 SRVY
814 - 817 RSRG

>PDOC00005 PS00005 PKC_PHOSPHO_SITE Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].
148 - 150 EAR
144 - 146 TGR
171 - 173 RSR
219 - 221 ESR
389 - 391 TRS
440 - 442 RGR
113 - 115 RGR
585 - 587 RLR
602 - 604 TRR
612 - 614 TRR
716 - 718 RGR
726 - 728 RGR
747 - 749 RGR
794 - 796 EAR
804 - 806 RGR
844 - 846 RLR
860 - 870 RSR
921 - 923 RGR
957 - 959 RSR
960 - 962 TRR
974 - 976 TRR
997 - 999 RLR
1062 - 1064 TRR
1018 - 1020 RGR
1031 - 1033 TRR
1119 - 1121 RGR
```

Sekundární databáze

- **Databáze funkčních nebo strukturních motivů** získaných srovnáním primárních dat (sekvencí)
- **PROSITE**, <http://www.expasy.org/prosite/>

Hits for all PROSITE (release 2020_05) motifs on sequence USERSEQ1 :

Round 2 hits in 1 sequence

USERSEQ1 (1122 aa)

```
MDVVTKLVAHRIVVFCVLAFLVVFECIWSNNTTIEWVQVSTEDLRTSLVSEIEHIGK
FTAKPLSTIGLAVYDSYITINDDTFEIQTALFLPWASTILQVWSYVSIHQGLAFSYIA
ESYFVAVFAESESSESDGQVQVQVQQIQAQHSREYKQGIQVYVQQVQAQGRITTEVY
DTLGEDEKITLQYVWLYSGQLVSGFFVWTLTEVLSLRSRSEEDIMCTKQVTVVREGSLM
DEFTTSDGCTPSEKESDSDGCTPSEKESDSDGCTPSEKESDSDGCTPSEKESDSDGCT
GATRIPIQAKAGVQLVVNI FLQFQFWYFWYVWQATPREDSDGATLIDKQETAQAKESKHK
VQFLPDAEIPKSLKLLWRRLLDCKQKFFVSDTFLVQVQVQVQVQVQVQVQVQVQV
WQVLEDPLKSLLEVDYFNPVAKWQVQVFLPQDQVFFVQVQVQVQVQVQVQVQVQV
PFTVQSLVANNQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
EVYVCTQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
QVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
PESSELPTAAGDISSQLPQRSFSAVAVLVIDAKQFFELDQVQVQVQVQVQVQVQV
WVESSEVWFQKQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
VQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
VDVETSDVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
QVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
```



Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not inter. For more information about how these graphical representations are constructed, go to <https://prosite.expasy.org/mydomains/>

Hits by profiles: [2 hits (by 2 distinct profiles) on 1 sequence]

Upper case represents match positions, lower case inset positions, and the "-" symbol represents deletions relative to the matching profile.



Sekundární databáze

- **Databáze funkčních** nebo **strukturních motivů** získaných srovnáváním primárních dat (sekvencí)
- **PROSITE**, <http://www.expasy.org/prosite/>

Hits for all PROSITE (release 2020_05) motifs on sequence USERSEQ1 :

found 2 hits in 1 sequence

USERSEQ1 (1122 aa)

```

MSITGTCASRSVYVCTALAPVVEELMLDNNITTRMVEVASSTEDLRTDLVDEIRSK
FTVAKLSTIIGLARVIGDVIINMDFPEKCIAPLIPVAIPIIDVQVSVIISDGLMFFVIA
ESITPVAVSSSSSSSSSDGRTWCTVYVQLVORLQNFVQDQDVTMTDVAAGQNTITPAPY
CISLQGEDEETLQVYVIVSISGRLVSLDFPWTGTEHSEKLRKISELQWPGCTVVEEGLI
DFFIIRSDICFRRESLKQKQIIRKCSISQVEIIRLFLVAFQVYEVVQVFLRVLKFRGQ
GQRIRIQGASEGKQLVTNIFIRFQRFVWQVQVRSKQKQATLQAGAQQQAKKQVQKQ
SGAFKASDIPKALANGQIDKCRGQVPSQDVTTLVQVVCACGLVLLSVMISKTESK
NGLTEPISLQLEEVTRFIRFQRFVWQVQVRSKQKQATLQAGAQQQAKKQVQKQ
KTVVSRIVAAHQKPKQSSSVLASVPVQVSTVQVQVCFQKQKQKQESTIETISINSHADME
PVEVYVYVGLIIRNDRKQYVETVYVQVETQVQVTLVQVQVQVQVQVQVQVQVQVQVQV
TCTQVYVTVTLLEPVSQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
FQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
FERRVFFQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
KLRSTVYVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
VQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
VQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
VQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
    
```

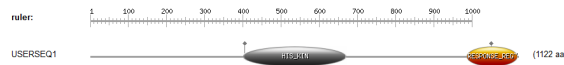
Legend:



Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not inter. For more information about how these graphical representations are constructed, go to <https://prosite.expasy.org/mydomains/>.

hits by profiles: [2 hits (by 2 distinct profiles) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.



Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáváním primárních dat (sekvencí)

□ PRINTS, <http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/>



PRINTS is a compilation of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of a PRINTS-PROFFPRINTS composite. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D space. Fingerprints can encode protein links and functionalities more flexibly and powerfully than can single motifs, full diagnostic potency deriving from the mutual context provided by motif neighbours. [References](#)

News:

- [SPINCE](#) - Search PRINTS and manual PRINTS
- [mpPRINTS](#) - Search PRINTS automatic application
- [PRINTS](#) - Search the improved [Invent](#) Pro family database

Direct PRINTS access:

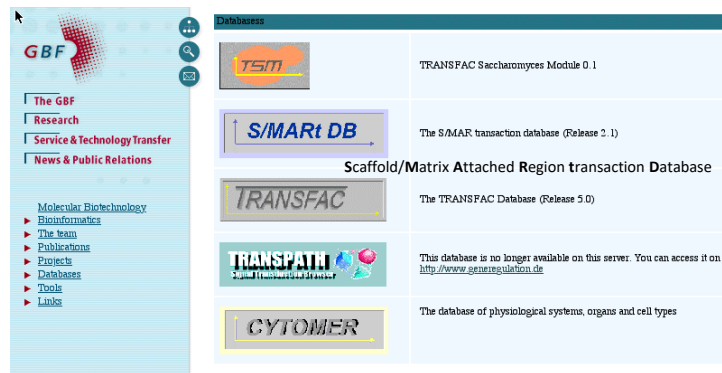
- [By sequence number](#)
- [By PRINTS code](#)
- [By UniProt code](#)
- [By name](#)
- [By domain](#)
- [By PDB](#)
- [By number of motifs](#)
- [By motif](#)
- [By expert taxonomy](#)

PRINTS search:

- Search PRINTS with [NEW InvenPRINTScan](#)
- [PRINTS](#)
- [CEPIScan](#)
- [MILScan](#)
- [FingerPRINTScan](#) binaries and source are available: printscan@bioinf.man.ac.uk

Sekundární databáze

- **TRANSFAC** <http://www.gene-regulation.com/>



The screenshot shows the TRANSFAC website interface. On the left is a navigation menu for GBF (Gene Bioinformatics Facility) with categories like 'The GBF', 'Research', 'Service & Technology Transfer', and 'News & Public Relations'. Below these are sub-categories: 'Molecular Biotechnology', 'Bioinformatics', 'The team', 'Publications', 'Projects', 'Databases', 'Tools', and 'Links'. The main content area is titled 'Databases' and lists several databases:

Database Name	Description
TSM	TRANSFAC Saccharomyces Module 0.1
S/MARt DB	The S/MAR transaction database (Release 2.1)
TRANSFAC	The TRANSFAC Database (Release 5.0)
TRANSFAC	Scaffold/Matrix Attached Region transaction Database
TRANSFAC	This database is no longer available on this server. You can access it on http://www.gene-regulation.com/
CYTOMER	The database of physiological systems, organs and cell types

46

CEITEC

S/MARt DB (saffold/matrix attached region transaction database). This database collects information about S/MARs and the nuclear matrix proteins that are supposed be involved in the interaction of these elements with the nuclear matrix. <http://transfac.gbf.de/SMARTDB/index.html>)

Strukturální databáze

- PDB <http://www.rcsb.org/pdb/>

The screenshot shows the PDB website homepage. At the top, there is a navigation bar with links for "DEPOSIT data", "DOWNLOAD files", "browse LINKS", "BETA TEST new features", and "BETA mirror files". Below this, there are sections for "Current Holdings" (19623 Structures, Last Update: 30-Dec-2002, PDB Statistics) and "Molecule of the Month: Cytochrome c". The main content area features a search bar with the text "Search the Archive" and "Enter a PDB ID or keyword". There are also links for "Query Tutorial", "Find a structure", and "PDB Mirrors". A "News" section is visible, dated "23-Dec-2002", with the headline "Happy Holidays from the PDB!". The footer of the page includes the CEITEC logo.

Strukturální databáze


- **PDB** <http://www.rcsb.org/pdb/>

Structure Explorer - 1P5Y

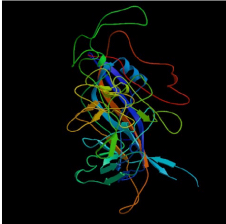
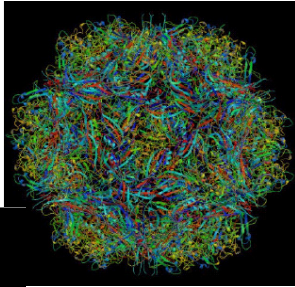
PDB
PROTEIN DATA BANK

Structure Explorer - 1P5Y

Title The Structures Of Heat Range Controlling Regions Of The Capsids Of Canine and Feline Parvoviruses and Mutants
Classification Virus/Viral Protein
Compound Mol. Bt. C. Molecular Coat Protein Vp2, Chain: A; Fragment: Sequence Database Residues 190-237; Engineered: Yes; Mutations: Yes
Exp. Method X-ray Diffraction

 **View Structure**

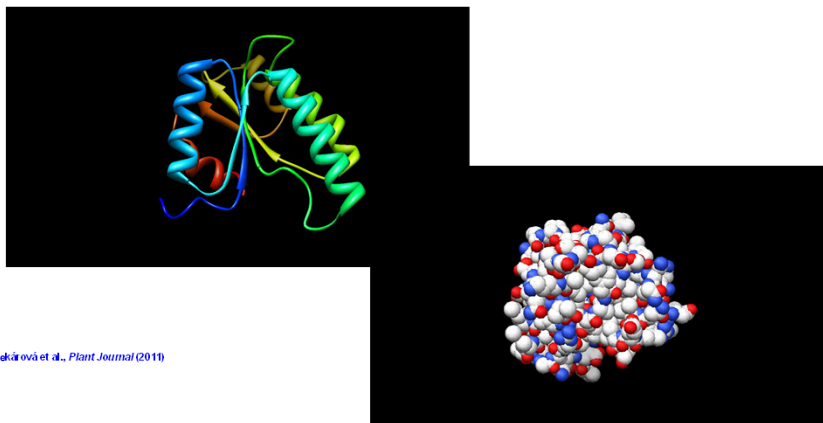
[Summary Information](#)
[View Structure](#)
[Download Display File](#)
[Structural Neighbors](#)
[Geometry](#)
[Other Sources](#)
[Sequence Details](#)



<http://www.rcsb.org/pdb/cgi/structure.cgi?job=graphics&pdbId=1P5Y;page=pid-173561064249344&bio=1&opt=show&size=500> 12/29/2003

Strukturální databáze

- PDB <http://www.rcsb.org/pdb/>



Pekárová et al., *Plant Journal* (2010)

Osnova

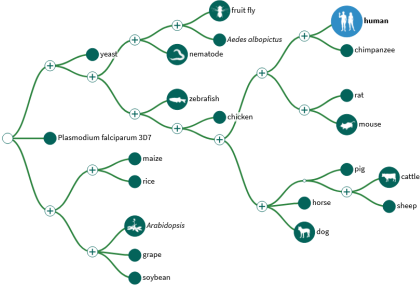
- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
 - Spektrum „on-line“ zdrojů
 - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
 - GENOMOVÉ zdroje

Genomové zdroje

□ [NCBI Genome Data Viewer](https://www.ncbi.nlm.nih.gov/genome/gdv/) <https://www.ncbi.nlm.nih.gov/genome/gdv/>

Genome Data Viewer GDV is a genome browser supporting the exploration and analysis of more than 920 eukaryotic RefSeq genome assemblies. 0

Select organism
Homo sapiens (human)



Homo sapiens (human) genome

Search in genome
Location, gene or phenotype

Examples: TP53, chr17:7687000-7688000, rs334, DNA repair

Assembly
GRCh38.p13

[Browse genome](#) [BLAST genome](#)

Assembly details

Name	GRCh38.p13
RefSeq accession	GCF_000001405.39
GenBank accession	GCA_000001405.28
Download via FTP	RefSeq, GenBank
Submitter	Genome Reference Consortium
Level	Chromosome
Category	Reference genome

Annotation details

Annotation Release	109
Release date	2020-08-17

GRCh38.p13
GRCh38.p12
GRCh38.p11
GRCh38.p10
GRCh38.p9
GRCh38.p8
GRCh38.p7
GRCh38.p6
GRCh38.p5
GRCh38.p4
GRCh38.p3
GRCh38.p2
GRCh38.p1
GRCh38.p0
GRCh38.p0.1
GRCh38.p0.2
GRCh38.p0.3
GRCh38.p0.4
GRCh38.p0.5
GRCh38.p0.6
GRCh38.p0.7
GRCh38.p0.8
GRCh38.p0.9
GRCh38.p0.10
GRCh38.p0.11
GRCh38.p0.12
GRCh38.p0.13
GRCh38.p0.14
GRCh38.p0.15
GRCh38.p0.16
GRCh38.p0.17
GRCh38.p0.18
GRCh38.p0.19
GRCh38.p0.20
GRCh38.p0.21
GRCh38.p0.22
GRCh38.p0.23
GRCh38.p0.24
GRCh38.p0.25
GRCh38.p0.26
GRCh38.p0.27
GRCh38.p0.28
GRCh38.p0.29
GRCh38.p0.30
GRCh38.p0.31
GRCh38.p0.32
GRCh38.p0.33
GRCh38.p0.34
GRCh38.p0.35
GRCh38.p0.36
GRCh38.p0.37
GRCh38.p0.38
GRCh38.p0.39
GRCh38.p0.40
GRCh38.p0.41
GRCh38.p0.42
GRCh38.p0.43
GRCh38.p0.44
GRCh38.p0.45
GRCh38.p0.46
GRCh38.p0.47
GRCh38.p0.48
GRCh38.p0.49
GRCh38.p0.50
GRCh38.p0.51
GRCh38.p0.52
GRCh38.p0.53
GRCh38.p0.54
GRCh38.p0.55
GRCh38.p0.56
GRCh38.p0.57
GRCh38.p0.58
GRCh38.p0.59
GRCh38.p0.60
GRCh38.p0.61
GRCh38.p0.62
GRCh38.p0.63
GRCh38.p0.64
GRCh38.p0.65
GRCh38.p0.66
GRCh38.p0.67
GRCh38.p0.68
GRCh38.p0.69
GRCh38.p0.70
GRCh38.p0.71
GRCh38.p0.72
GRCh38.p0.73
GRCh38.p0.74
GRCh38.p0.75
GRCh38.p0.76
GRCh38.p0.77
GRCh38.p0.78
GRCh38.p0.79
GRCh38.p0.80
GRCh38.p0.81
GRCh38.p0.82
GRCh38.p0.83
GRCh38.p0.84
GRCh38.p0.85
GRCh38.p0.86
GRCh38.p0.87
GRCh38.p0.88
GRCh38.p0.89
GRCh38.p0.90
GRCh38.p0.91
GRCh38.p0.92
GRCh38.p0.93
GRCh38.p0.94
GRCh38.p0.95
GRCh38.p0.96
GRCh38.p0.97
GRCh38.p0.98
GRCh38.p0.99
GRCh38.p0.100

Genomové zdroje

□ **Genome Browser Gateway** <https://genome.ucsc.edu/>

The UCSC Genome Browser was created by the Genome Biotechnology Division of UC Santa Cruz. Software Copyright (c) The Regents of the University of California. All rights reserved.

chr genome assembly
Chr: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Genome: hg19 hg18 hg17 hg16 hg15 hg14 hg13 hg12 hg11 hg10 hg9 hg8 hg7 hg6 hg5 hg4 hg3 hg2 hg1
Assembly: hg19 hg18 hg17 hg16 hg15 hg14 hg13 hg12 hg11 hg10 hg9 hg8 hg7 hg6 hg5 hg4 hg3 hg2 hg1

Human Genome Browser - hg19 assembly (sequences)
This is February 2009 human reference sequence (hg19) was produced by the Genome Reference Consortium. For more information about this assembly, see [hg19](#) in the NCBI Assembly database.

Sample position queries
A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the [Query's Guide](#) for more information.

Request	Genome Browser Response
chr7	Displays all of chromosome 7
chr3:45000-12	Displays all of the region of chr3:45000-12
2p13	Displays region for band p13 on chr 2
chr3:100000	Displays 100 bases of chr 3, starting from 0-prim location
chr3 100000-2000	Displays a region of chr3 that spans 2000 bases, starting with position 100000
RR10001 RR40115	Displays region between genome landmarks, such as the STS markers RR10001 and RR40115, or chromosome bands: 15q11 to 15q13, or SNPs rs104522 and rs1800370. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, miRNAs, etc.
rs104522:rs1800370	
D1003046	Displays region around STS marker D1003046 from the chromosome's original map. It includes 100,000 bases on each side as well
A020414	Displays region of EST with GenBank accession A020414 in a BRCA1 cancer gene on chr 17
AC008103	Displays region of clone with GenBank accession AC008103
A020311	Displays region of mRNA with GenBank accession number A020311
PRNP	Displays region of genome with HUGO Gene Nomenclature Committee identifier PRNP
MIM_211414	Displays the region of genome with MIM identifier MIM_211414
NP_059110	Displays the region of genome with protein accession number NP_059110
proteinogen mRNA	Lists transcribed products, but not cDNAs
homobox caudal	Lists mRNAs for caudal homobox genes
zinc finger	Lists many zinc finger mRNAs
haptoglobin	Lists only haptoglobin cDNA clones
huntingtin	Lists candidate genes associated with Huntington's Disease
zab19	Lists mRNAs deposited by scientist named Zaher
Evans_J_E	Lists mRNAs deposited by co-author J.E. Evans

Genomové zdroje

□ Human Genome Browser <http://genome.ucsc.edu/cgi-bin/hgGateway>

The screenshot displays the UCSC Genome Browser interface for the Human Feb. 2009 (GRCh37/hg19) Assembly. The browser shows a search bar at the top with the coordinates chr11:5,246,895-5,245,301 (1,000 bp). Below the search bar, there are several tracks including the RefSeq track, which is highlighted with a green arrow. The RefSeq track shows gene models for the region. Other tracks include the UCSC Genes track, the RepeatMasker track, and the BLAT track. The bottom of the browser shows a list of tracks that can be toggled on or off, such as the Chromosome Band, BSL Markers, and BSL Clones tracks.

Genomové zdroje

Human Genome Browser <http://genome.ucsc.edu/cgi-bin/hgGateway>

The image shows a screenshot of the Human Genome Browser interface. At the top, there is a navigation bar with tabs for 'Genes', 'Genes (Detailed)', 'Genes (Detailed) (1)', and 'Genes (Detailed) (2)'. Below this, the main content area displays information for the HBB gene. The 'Description and Page Info' section includes a description of the gene's function in hemoglobin synthesis. The 'Page Info' section lists various genomic data sources like Genomic Sequence, Gene Sorter, and COSM. A green arrow points to the 'Sequence and Links to Tools and Databases' section. The 'Comments and Description Text from UniProtKB' section contains detailed scientific information about the HBB protein, including its function, structure, and associated diseases like beta-thalassemia.

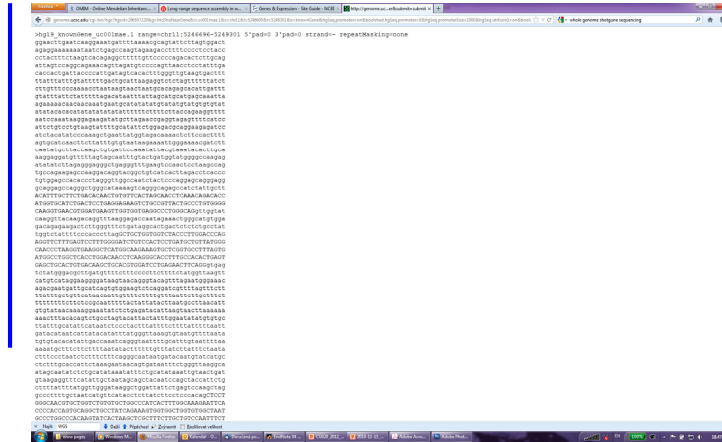
Genomové zdroje

□ **Human Genome Browser** <http://genome.ucsc.edu/cgi-bin/hgGateway>

The screenshot shows the 'Get Genomic Sequence Near Gene' interface of the Human Genome Browser. The page title is 'Genomic Sequence Near Gene'. Below the title, there is a note: 'Note: if you would prefer to get DNA for more than one feature of this track at a time, try the [Table Browser](#) using the output format sequence.' The main section is 'Sequence Retrieval Region Options:' and contains several checkboxes and input fields: 'Flanking upstream by 1000 bases' (checked), '5' UTR Exons' (checked), 'CDS Exons' (checked), '3' UTR Exons' (checked), 'Introns' (checked), 'Downstream by 1000 bases' (checked), and 'One FASTA record per region (exon, intron, etc.) with 5' extra bases upstream (5') and 3' extra downstream (3')' (checked). Below this is a note: 'Note: If a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.' The 'Sequence Formatting Options:' section includes: 'Exons in upper case, everything else in lower case' (checked), 'CDS in upper case, UTR in lower case' (checked), 'All upper case' (unchecked), 'All lower case' (unchecked), and 'Mask repeats: # to lower case | to N' (unchecked). The page is mostly blank with a large yellow area where the sequence would be displayed.

Genomové zdroje

Human Genome Browser <http://genome.ucsc.edu/cgi-bin/hgGateway>



Genomové zdroje

- The Arabidopsis Information Resource (TAIR) <http://www.arabidopsis.org>



Genomové zdroje

- The Arabidopsis Information Resource (TAIR) <http://www.arabidopsis.org>



The Arabidopsis Information Resource

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana*. Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Gene product function data is updated every two weeks from the latest published research literature and community data submissions. Gene structures are updated 1-2 times per year using computational and manual methods as well as community submissions of new and updated genes. TAIR also provides extensive linkouts from our data pages to other Arabidopsis resources.

The Arabidopsis Biological Resource Center at The Ohio State University collects, reproduces, preserves and distributes seed and DNA resources of *Arabidopsis thaliana* and related species. Stock information and ordering for the ABRC are fully integrated into TAIR.

Breaking News

Data Updates Suspended
[October 15, 2006]
Some TAIR data updates, including loading of new ABRC stocks, will be suspended from Oct 20-Nov 17 while we move our servers.

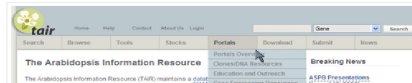
New Phenotype Search Option
[October 15, 2006]
Search for genes, germplasm and polymorphisms using associated phenotype, and see improved phenotype data display in results and detail pages.

The NEW arabidopsis.org

We've added new dropdown headers and left navigation bars and reorganized our web pages to make it easier to locate information and resources in TAIR. Please contact us if you experience any problems with our new site.

ASPB Presentations
[August 15, 2006]
Following heavy demand, the TAIR workshop presentations given at the ASPB meeting in Boston have been made available from the TAIR website for download.

AHP2 @ TAIR



Osnova

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
 - Spektrum „on-line“ zdrojů
 - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
 - GENOMOVÉ zdroje
- Analytické nástroje
 - Vyhledávání homologí

Analytické nástroje

□ Globální vs. lokální přiřazení

```
Globální přiřazení
SLAV-----APATNIK-----PIQNYR-I-----AKSETQRYMVE
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRREHVRATAKSETQRYMVE

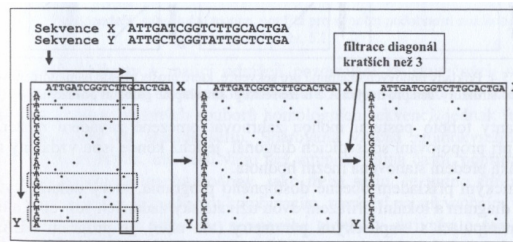
Lokální přiřazení
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRREHVRATAKSETQRYMVE
-----NAPATNIKSECVRA-PIQNYRREHVR-----
```

Cvrčková, Úvod do praktické bioinformatiky

- **Globální přiřazení** pouze u sekvencí, které jsou si **podobné a podobné délky** (za cenu vnášení mezer do jedné nebo obou sekvencí)
- Globální přiřazení se používá především v případě **mnohačetného přiřazování** (CLUSTALW, viz dále)
- **Lokální přiřazení** umožní identifikaci a srovnání i v případě porovnávání pouze **úseků sekvencí** s významnou mírou podobnosti, např. i při záměně pořadí proteinových domén během evoluce

Analytické nástroje

- Volba správného typu přiřazení pomocí bodového diagramu (dotplot)

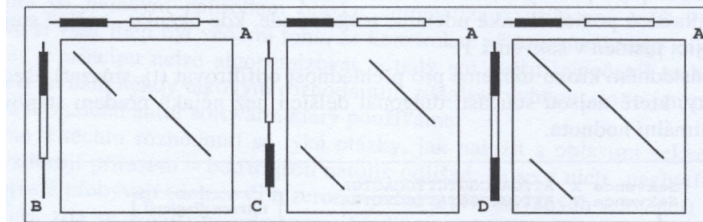


Cvrčková, Úvod do praktické bioinformatiky

- vynesení sekvencí proti sobě
- identifikace shody v okně o dané velikosti (např. 2 bp)
- „odfiltrování“ diagonál o délce menší než je mezní hodnota (threshold)

Analytické nástroje

- příklady srovnání sekvencí pomocí bodového diagramu



Cvrčková, Úvod do praktické bioinformatiky

- **globálně** lze srovnávat **pouze** sekvence A, B
- ostatní sekvence prošly během evoluce **záměnou domén** a je nutné je porovnávat **lokálně**
- **bodový diagram** lze získat pomocí srovnávání programem BLAST2 (viz dále)

Analytické nástroje

- **BLAST** <http://ncbi.nlm.nih.gov/BLAST/>

The screenshot shows the NCBI Nucleotide BLAST search page. At the top, the NCBI logo is on the left, and the text "nucleotide-nucleotide BLAST" is on the right. Below this, there are four tabs: "Nucleotide", "Protein", "Translations", and "Retrieve results for an RID". The "Nucleotide" tab is selected. In the center, there is a text input field containing the following DNA sequence: "aacacacacg acacacacat cattatcaco atcgttttgg ggcgatgttg tgtggttoca ggtatataat ataattaatt tattccacat gagatatgat atgatatact atgtatTTTT ttattgttaa acctttaata taacaagaac tacaaaaaat gaaaa". Below the input field, there are two buttons: "Set subsequence" and "Choose database". The "Set subsequence" button has "From:" and "To:" input fields. The "Choose database" button has a dropdown menu showing "nr". At the bottom, there are three buttons: "BLAST!" (highlighted in blue), "Reset query", and "Reset all".

BLAST

Basic Local Alignment Search Tool

>gi|5016088|ref|NM_001101.2| Length = 1793 E= expectancy value actin, beta (ACTB), mRNA

Score = 1110 bits (560), Expect = 0.0
Identities = 965/1100 (87%)
Strand = Plus / Plus

Query: 156 gtcgacaacggctctggcatgtgcaaggccggatttgcggagacgatgctccccggccc 215
Sbjct: 101 gtcgacaacggctccggcatgtgcaaggccggcttcgggggagacgatgccccggggcc 160

Query: 216 gtcttcccacgatgtgggaogtcccogtcaccaggggtgtgatggctggcagggccag 275
Sbjct: 161 gtcttcccctccatctgtgggggccccaggaaccagggogtgatgggtggcatgggtcag 220

Query: 276 aaggactcgtacgtgggtgatgaggccagagcaagcgtggtatcctcaccctgaagtac 335
Sbjct: 221 aaggatcctatgtggggacgaggccccagagcaagagaggaatcctcaccctgaagtac 280

Query: 336 cccattgagcacggtatcgtgaccaactgggacgatggagaagatctggcaccacacc 395
Sbjct: 281 cccatcgagcacggcatcgtcaccactgggacgatggagaagatctggcaccacacc 340

ds..S=1213 E=0.0
>=200
250 1500

- „expectancy value“ udává předpokládaný počet sekvencí se stejnou nebo lepší podobností při vyhledávání ve stejně velké databázi složené z náhodných sekvencí
- výsledek udává frakci totožných a u proteinů i podobných pozic, příp. počet vložených mezer

Primární databáze

The screenshot displays a GenBank record for the gene NP_059797.1. The main view shows a genomic map with a scale from 145,400 to 147,600. A red bar represents the gene, and a green arrow points to its start. A popup window provides the following details:

- NP_059797.1**
- NP_059797.1: two-component VirA-like sensor kinase
- total range: NC_002377.1 (145,694..148,183)
- total length: 2,490
- strand: plus
- protein product length: 829
- Links & Tools**
- GenBank View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1](#)
- FASTA View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1](#)
- BLAST Genomic: [NC_002377.1 \(145,694..148,183\)](#)
- Graphical View: [NP_059797.1](#)
- BLAST Protein: [NP_059797.1](#)
- BLINK Results: [NP_059797.1](#)

Below the popup, there are sections for **Bibliography** and **Related articles in PubMed**.

66

CEITEC

BLINK is a link to the pre-computed BLAST search results for the respective sequence (see the next slide).

BLAST

Basic Local Alignment Search Tool

Pre-computed BLAST results for: [gi|16119781|ref|NP_396486.1](#) two component sensor kinase [Agrobacterium tumefaciens str. C58]

Matching gis: [15162425.20141871.101960](#)

Total (score > 100) : 147068 hits in 148754 proteins in 6309 species

Selected: 147068 hits in 146754 proteins in 6309 species Filter: Min Score: 100

Other views (Reports): [Taxonomy report](#) | [Multiple Alignment](#) | [Blast](#)

[Reset all filters](#)

Choose Display Options

1203 Archaea 138095 Bacteria 13 Metazoa 1348 Fungi 564 Plants 6 Viruses 5676 The Others [reset selection](#)

Results: 1 - 100 [Next Page](#) [Last](#)

Rank	% hit	Score	Accession	Length	Protein Description
1	100	41.66	AAK90927	833	two component sensor kinase [Agrobacterium tumefaciens str. C58]
2	100	41.66	F18549	833	hecton: full-wide host range vira protein Short-WIR vira
3	100	41.66	AAK79282	833	vira [Pisamid pTCS8]
4	100	41.59	NP_053330	833	hypothetical protein pTI-SAKUPA_p142 [Agrobacterium tumefaciens]
5	100	41.59	AAK17465	833	tiorf140 [Agrobacterium tumefaciens]
6	100	41.53	AAK93269	833	vira [Pisamid T1]
7	100	41.53	gi1737127	833	vira protein
8	100	41.53	CAA14727	833	91.2 kDa protein [Agrobacterium tumefaciens]
9	100	39.00	CAA35269	829	vira [Agrobacterium rhizogenes]
10	100	37.18	gi1227240	849	vira gene
11	100	33.48	AAK8643	829	vira [Pisamid T1]

BLAST

Specializované verze

- V současnosti existuje celá řada specializovaných verzí programu [BLAST](#)
 - vyhledávání podle zdroje (organismu) sekvencí, např. známých genomů [mikroorganismů](#)
 - **BLASTP**
 - vyhledávání podobnosti k [proteinu](#) v [databázi proteinových sekvencí](#)
 - **BLASTN**
 - vyhledávání podobnosti k [nukleotidové sekvenci](#) v [databázi nukleotidových sekvencí](#)
 - další varianty jako např. [MEGABLAST](#) pro identifikaci totožných nebo velice podobných sekvencí (vyhledává [dlouhé podobné úseky nukl. sekvencí](#))
 - **BLASTX**
 - vyhledávání [podobnosti nukleotidové sekvence](#) přeložené do sekvence [aa](#) v [proteinové databázi](#)

BLAST

Specializované verze

- V současnosti existuje celá řada specializovaných verzí programu BLAST
 - **TBLASTN**
 - vyhledávání podobnosti **proteinové sekvence** v nukleotidové databázi přeložené do sekvence aa
 - **TBLASTX**
 - vyhledávání k **sekvenci nukleotidů přeložené** do sekvence aa v **databázi nukleotidových sekvencí přeložených** do sekvence aa

BLAST

Specializované verze

- V současnosti existuje celá řada specializovaných verzí programu BLAST
 - **PSI-BLAST** (Position-Specific Iterated BLAST)
 - Prvním krokem je standardní BLAST, při kterém PSI-BLAST identifikuje skupinu podobných sekvencí s E hodnotou lepší než minimální hodnota (standardně 0,005)
 - PSI-BLAST vytváří pro každé přiřazení tzv. **PSSM** (Position Specific Substitution Matrix)
 - PSSM matice zohledňuje výskyt jedné aminokyseliny ve stejné pozici se zvýšenou frekvencí u sekvencí identifikovaných jako podobné v prvním kole pomocí BLAST, což může znamenat funkční konzervovanost

BLAST

Specializované verze

- V současnosti existuje celá řada specializovaných verzí programu BLAST
 - **PHI-BLAST** (Pattern-Hit Initiated **BLAST**)
 - Určen k identifikaci specifické sekvence, např. motivu (pattern) v sekvenci podobných proteinových sekvencí
 - Sekvenci motivu je třeba vložit pomocí **speciálního syntaxu**
 - [LVIMF] znamená buď Leu, Val, Ile, Met nebo Phe
 - - je oddělovník (neznačená nic)
 - x(5) znamená 5 jakýchkoliv aminokyselin
 - x(3, 5) znamená 3 až 5 jakýchkoliv aminokyselin

BLAST

Specializované verze

□ Příklad vyhledávání pomocí PHI-BLAST

```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase  
MSHIQIPPGLTPELLQGYTVBVLRRQQPPDLVEFAVEYFTRLREARAPASVLPAAATPRQSLGHPPPPEPGPDR  
VADAKGDSSESEDEDELEVVPSPRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCLQBACKDILLF  
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSRGELA  
LMYNTPRAAITVA TSEGSLWGLDRVTFRRIIVKNNAKRKMFPESPIESVPLLKSLVSERMKIVDVIgek  
IYKDGERIITQGEKADSFYIIESGSEVSLIIRSRTKSNKDGNGQVEIARCHKQYFGELALVTKPRAAS  
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNISHYEEQLVKMFGSSVDLGNLGG  
  
[LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV].
```


Osnova

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
 - Spektrum „on-line“ zdrojů
 - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
 - GENOMOVÉ zdroje
- Analytické nástroje
 - Vyhledávání homologí
 - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restričních míst....

Analytické nástroje

<https://blog.addgene.org/free-online-molecular-biology-tools>

Early Career Researcher Toolbox: Free Online Molecular Biology Tools

By Beth Kenkel



Beth Kenkel
September 12, 2023

Share this article



Primer design. Plasmid mapping. DNA sequence analysis. We all have our favorite tools for tackling these particular tasks, but they tend to be scattered about the internet. To help you keep your virtual molecular biology toolbox organized, today's post features a list of free online molecular biology tools all in one place.

Plasmid mapping

These tools are for viewing, editing or making plasmid maps, but can also analyze and annotate any DNA sequence.

- **SnapGene Viewer**: The free SnapGene Viewer is great for looking at plasmid maps and viewing sequencing traces, while the paid version provides more tools for plasmid mapping and design (Figure 1).
- **Benchling**: While you might think of Benchling as an electronic lab notebook, it also has a suite of molecular biology tools and can make plasmid maps. Free for academic users.
- **Serial Cloner**: Free desktop-based software for plasmid design and mapping.
- **ApE (A plasmid Editor)**: A free, donation-based plasmid analysis tool including editing, annotating, creating maps, and more. This tool is maintained by M. Wayne Davis from the University of Utah.

SnapGene

<https://www.snapgene.com/snapgene-viewer/download>



<https://www.youtube.com/watch?v=0sQh2s182WQ>

Osnova

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
 - Spektrum „on-line“ zdrojů
 - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
 - GENOMOVÉ zdroje
- Analytické nástroje
 - Vyhledávání homologí
 - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restričních míst...
 - Další [www genomové nástroje](#)

Další WWW zdroje

- TIGR (The Institute for Genomic Research), <http://www.tigr.org/software/>
 - Recently part of the J. Craig Venter Institute

The screenshot shows the NCBI Gene database entry for PHACTR8. Key information includes:

- Gene:** PHACTR8 phosphatase and actin regulator 4 [Homo sapiens]
- Official Symbol:** PHACTR8
- Official Full Name:** phosphatase and actin regulator 4
- Primary Source:** HGSC:2573
- Location:** RP11-422C24_A1
- Genomic context:** Chromosome 1, NC_000011.0 (2869693-2869891)
- Genomic regions, transcripts, and products:** Includes a diagram of the gene structure on the chromosome and a list of transcripts and products.

Další WWW zdroje

- Online Mendelian Inheritance in Man (OMIM) <http://www.omim.org/>



Shrnutí

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
 - Spektrum „on-line“ zdrojů
 - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
 - GENOMOVÉ zdroje
- Analytické nástroje
 - Vyhledávání homologií
 - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restričních míst....
 - Další [www.genomové nástroje](#)

Diskuse